

CPE345: Modeling and Simulation

Lecture 10

Today's topic

- Input modeling for simulation
- Project report guidelines

Input Modeling

- The simulation implements a model for the real system, driven by some input data (usually computer generated)
 - Developing a correct model for the input data is crucial for obtaining significant output results
 - How to develop a model for data?
 - Should you collect data and fit it to a known distribution?
 - Is there enough data and time available?
 - Should you rely on references that have already proposed and validated a model for similar input data?
 - Is this model largely adopted in the community?

Data Modeling Steps

- Collect data
- Identify a *pdf* to model the observed data
- Determine the parameters of the distribution
- Evaluate goodness of fit (χ^2 or Kolmogorov-Smirnov)
 - Possibly repeat from step 2 (if the goodness of fit tests fail).

Collect data

- Data collection – very important but hard to achieve
 - Data may be too scarce or too abundant
 - Simulation results (output will depend on your input accuracy): GIGO (“garbage in, garbage out”)
- Data collection – some suggestions
 - Planning: preobserving session → watch for unusual circumstances and plan how to handle them
 - Analyze data as they are being collected – this may determine if the data is adequate
 - Look for the homogeneity in data sets (same means, and distributions) and combine similar data sets
 - Data censoring? Are all the quantities observed in their entirety, or the observation process is artificially truncated?

Collect data – cont.

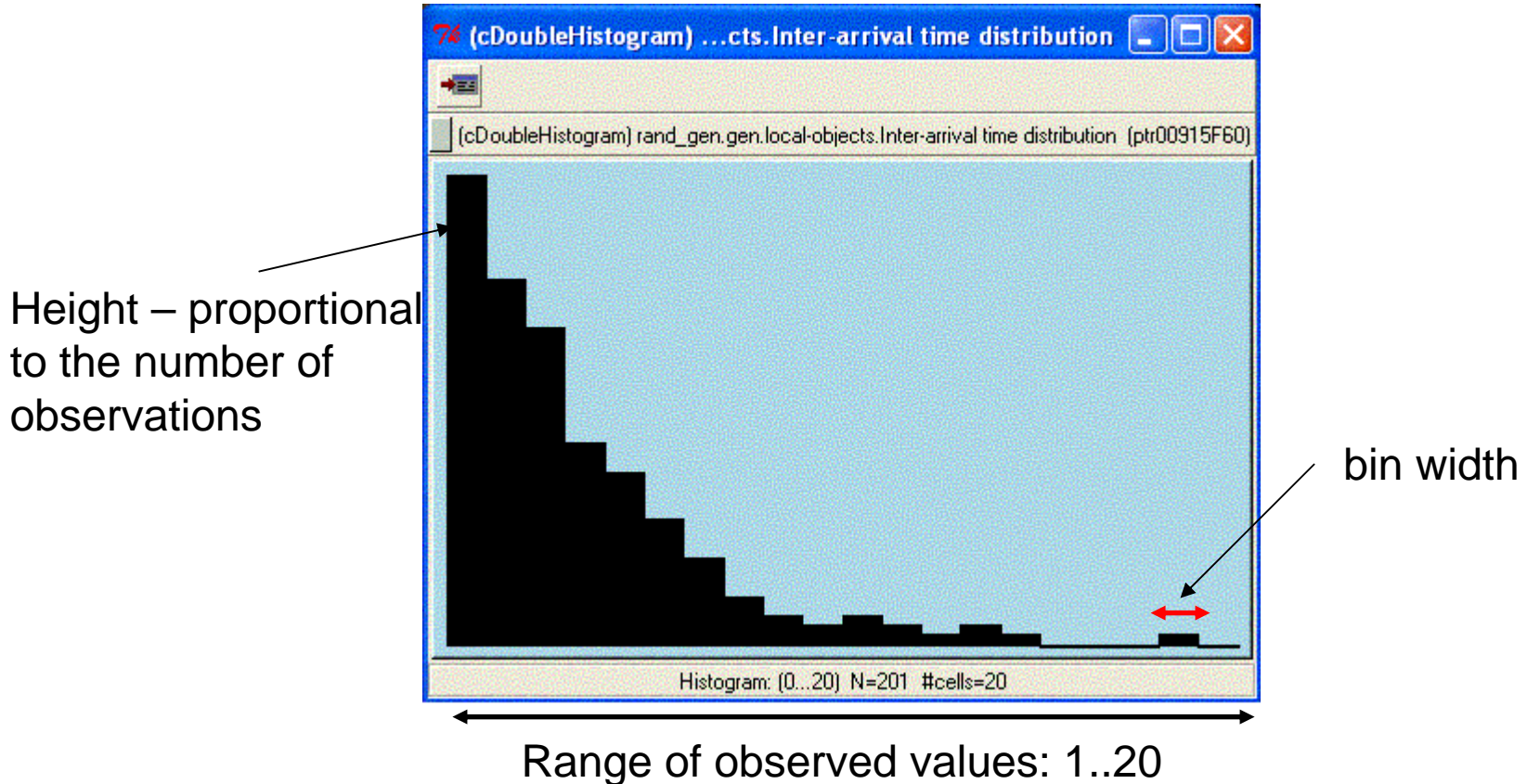
- Search for relationship between variables – use scatter plot to determine these relationships
- Look for correlation in data
- Input data \neq performance data \neq output data
 - Collect input data

Identify *pdf*

- How to identify distributions?
 - Histograms
 - Selecting a family of distributions
 - Quantile-quantile plots

Histograms

- We used them last lecture as OMNET++ output to show the simulated distribution



Histograms – cont.

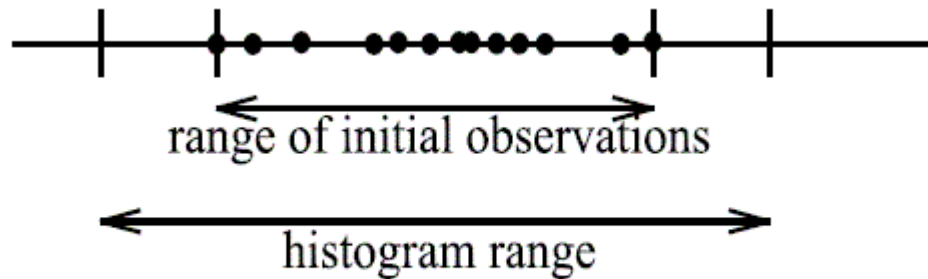
- Selecting the number of bins in a histogram for a given range
 - Depends on the number of samples and on the scatter (dispersion of data)
 - If too many, the bins are narrow, and the histogram may look ragged (will not smooth the data)
 - If too little, the bins are large, the histogram may be coarse
 - Some references suggest $\text{number of bins} \sim \sqrt{\text{sample size}}$
- The range can also be estimated by using a set of initial observations
 - Example in OMNET++: create a histogram with 20 cells, and automatic range selection:
`cDoubleHistogram histogram("histogram", 20);`
`histogram.setRangeAuto(100, 1.5);`

Number of initial observations

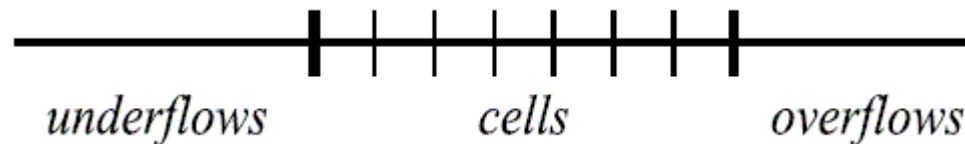
range extension factor

Automatic range estimation

- Range extension factor:



- Random numbers generated may still fall outside the selected range:

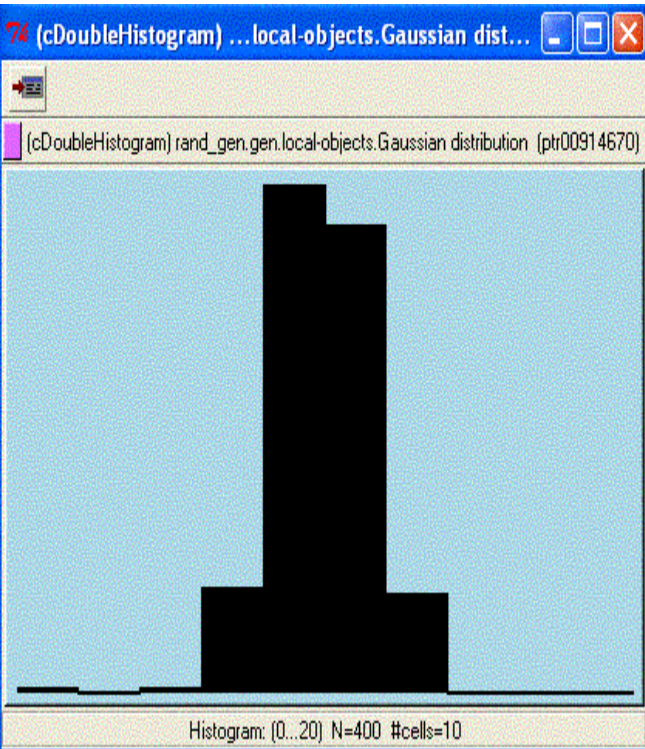


- The number of underflows and overflows is returned in OMNET++ by: `underflowCell()`, and `overflowCell()` functions

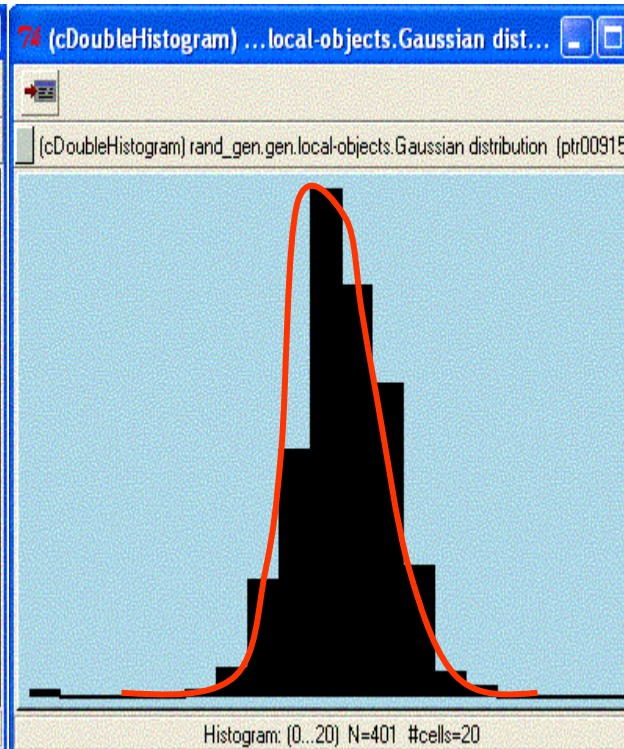
The effect of selecting the number of bins

$N = 400 \rightarrow \# \text{ bins} \cong 20$

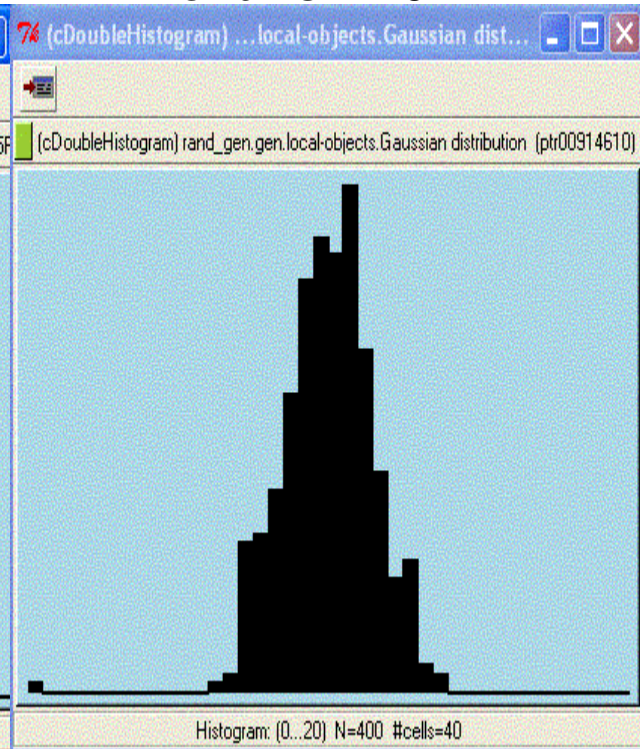
#bins = 10



#of bins = 20



of bins = 40



Selecting the family of the distribution

- Select from the *pdfs* that can be proposed as theoretical models
- How to select the distribution
 - Is it discrete or continuous?
 - Are the random variables (observations) bounded or there is no natural bound?
 - Can you infer the distribution from your knowledge about the process that generates input values? Examples:
 - Gaussian – sum of a large number of independent r.v.
 - Erlang – sum of exponential r.v.
 - Binomials: models the number of successes in n trials
 - Geometric: the number of trials until success
 - Poisson: number of independent events that occur in a fixed amount of time or space
 - Exponential: time between independent events, or a process time which is memoryless
 - Uniform: complete uncertainty – all outcomes are equally likely

Quantile-Quantile plots

- Verify your distribution selection using (q-q) plots
 - If X is a r.v., the q -quantile of X is that value γ , such that

$$F(\gamma) = P(X \leq \gamma) = q, \quad 0 < q < 1 \Rightarrow \gamma = F^{-1}(q)$$

- If we have a sample of data from X , $\{x_i, i = 1, 2, \dots, n\}$, the ordered sequence $\{y_j, j = 1, 2, \dots, n\}$ ($j=1$ is smallest, $j=n$ is largest), approximates

$$y_j \cong F^{-1}\left(\frac{j-1/2}{n}\right)$$

- Idea: plot $\{y_j, j = 1, 2, \dots, n\}$, versus the values $F^{-1}\left(\frac{j-1/2}{n}\right)$ for the selected distribution
- If the distribution match \rightarrow we will get a straight line with slope 1
- Ow, the plot will not be linear
- Study example 9.4 (chapter 9, page 334 in your textbook)

Parameter Estimation

- Mean and variance – maximum likelihood estimators
 - n samples

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

$$S^2 = \frac{\sum_{i=1}^n X_i^2 - n\bar{X}^2}{n-1}$$

- If the discrete data is grouped in k frequency distribution classes:

$$\bar{X} = \frac{\sum_{j=1}^k f_j X_j}{n}$$

$$S^2 = \frac{\sum_{j=1}^k f_j X_j^2 - n\bar{X}^2}{n-1}$$

Parameter Estimation – Cont.

- For continuous data, if data have been placed in class intervals
 - c frequency classes, f_j = observed frequency in the j^{th} class interval, m_j = midpoints of the frequency classes

$$\bar{X} \cong \frac{\sum_{j=1}^c f_j m_j}{n}$$

$$S^2 \cong \frac{\sum_{j=1}^c f_j m_j^2 - n\bar{X}^2}{n-1}$$

Most common distributions

- Poisson

- Parameter: α , estimate (from raw data): $\hat{\alpha} = \bar{X}$

- Exponential

- Parameter: μ , estimate (from raw data): $\bar{\mu} = \frac{1}{\bar{X}}$

- Gaussian

- Parameters: μ, σ^2 , estimates (from raw data):
 $\bar{\mu} = \bar{X}$
 $\hat{\sigma}^2 = S^2$

Goodness of fit

- We have already talked about hypothesis testing and have introduced the Kolmogorov-Smirnov and chi-square (χ^2) tests
- This tests will not give a definite answer with respect to the real distribution – should only guide the selection
 - Sample size may have significant effect on results
 - Small number of data points → no candidates rejected
 - Large number of data points → all candidates may be rejected

Goodness of fit tests

- χ^2 (chi square)
 - Compares the histogram of data with the shape of a candidate density function
 - Valid for large sample sizes
 - Assumes parameters estimated using the maximum likelihood formulas previously discussed
- χ^2 (chi square) with equal probabilities
 - If we assume that the distribution is continuous, the class intervals should be selected equal probability rather than equal width
- Kolmogorov-Smirnov
 - Based on examining a q-q plot
 - Very useful when you don't want to estimate any parameters from data, and when the sample size is small
 - Study example 9.15

χ^2 (chi square) with equal probabilities: an example

- Example 9.14: Exponential distribution
 - H0: distribution is exponential
 - H1: distribution is not exponential
 - First question: how many class intervals?
 - Recall χ^2 test condition:

$$E_i \geq 5 \Rightarrow np_i \geq 5 \Rightarrow \frac{n}{k} \geq 5 \Rightarrow k \leq \frac{n}{5}$$

- Unequal intervals \rightarrow determine the endpoints for the intervals
 - Requirement: equal probability for all intervals

χ^2 (chi square) with equal probabilities: an example – cont.

- For $n=50$ samples $\rightarrow k \leq 10$
- Select $k = 8 \rightarrow p = 0.125 \rightarrow E_i = n \cdot p = 6.25$
- Denote the end points with a_i . Then:

$$F(a_i) = 1 - e^{-\lambda a_i} = ip$$

exponential distribution

cumulative area from 0 to a_i

$$e^{-\lambda a_i} = 1 - ip \Rightarrow a_i = -\frac{1}{\lambda} \ln(1 - ip), \quad i = 0, 1, 2, \dots, k$$

$$\begin{cases} a_0 = 0 \\ a_8 = \infty \end{cases}$$

Example: tabulated results

- Estimated parameter: $\hat{\lambda} = 0.084$

Class interval	Observed freq. O_i	Expected freq. E_i	$\frac{(O_i - E_i)^2}{E_i}$
[0, 1.590)	19	6.25	26.01
[1.590, 3.425)	10	6.25	2.25
[3.425, 5.595)	3	6.25	0.81
[5.595, 8.252)	6	6.25	0.01
[8.252, 11.677)	1	6.25	4.41
[11.677, 16.503)	1	6.25	4.41
[16.503, 24.755)	4	6.25	0.81
[24.755, ∞)	6	6.25	0.01
Total	50	50	39.6

The statistic $\chi_0^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$ is chi-square distributed with **$k-s-1$ degrees of freedom**

- s = the number of parameters estimated for the hypothesized distribution

For $\alpha = 0.05 \rightarrow$ critical value is 12.6 \rightarrow the H_0 hypothesis is rejected

Some observations about goodness of fit tests

- The outcome of the tests depends on the significance level. Common values are 0.01, 0.05, and 0.10
- Sometimes you reject a hypothesis for some value, but the test is valid for another one
- How to choose the significance level?
 - The significance level = probability of falsely rejecting H_0
 - Can compute a *p-value* = the significance level for which one would just reject H_0 (for the given value of the test statistic)
 - Larger value for *p-value* is better – indicates a good fit
 - With respect to testing distributions:
 - Test many possible distributions
 - Select the one with the largest p-value

Selecting input models without data

- When no data is available
 - Use references that describes models for the input data
 - Get expert opinion – help to characterize the inputs
 - Use physical limitations to bound the problem (e.g. computer data entry cannot be faster than a person can type)
 - Exploit the nature of the process – choose a distribution that is most closely related to the underlying input process
 - In general, uniform, triangular and beta distributions are used when nothing else is known

Multivariate and Time Series Input Models

- In the previous discussion we have assumed independent processes
- Sometimes multiple inputs to the same process are related to each other
 - Multivariate input models – fixed finite number of random variables
 - E.g. for two normally distributed random variable, their dependence may be modeled as a bivariate normal distribution
 - Time series input models of a sequence of related r.v.
- Two measure of dependence between random variables:
 - Covariance
 - Correlation

Measures of dependence

- If X_1 and X_2 are two r.v. with mean μ_i and variance σ_i^2

- Covariance:

$$\text{cov}(X_1, X_2) = E[(X - \mu_1)(X - \mu_2)] = E(X_1 X_2) - \mu_1 \mu_2$$

- Correlation:

$$\rho = \text{corr}(X_1, X_2) = \frac{\text{cov}(X_1 X_2)}{\sigma_1 \sigma_2}$$

- Properties: $-1 \leq \rho \leq 1$

- Correlated if $|\rho| \geq 0.33$

- Uncorrelated if $|\rho| \leq 1$

Project: final report guidelines

- The report should contain
 - **Introduction** – what problem you studied
 - **Assumptions** – system model and justifications for the model
 - **Input data modeling** – collection, references, inferences about system?
 - **Simulation program**
 - **Code listing**
 - **Network picture**
 - **Simulation results** (usually output graphs and statistical averages, e.g., delays, throughput, etc)
 - **Validation/verification** – is the model and the simulation meaningful – arguments
 - **Conclusions and recommendations** – what did you observe, which of the solutions is better...
 - **References**

Homework

- Not to be handed in: problem 16, chapter 9, page 362